Ref 4

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-132559

(43)Date of publication of application : 12.05.2000

(51)Int.Cl. G06F 17/30

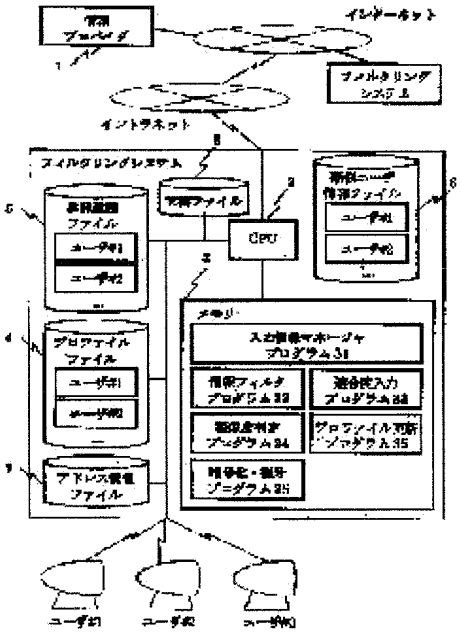
(21)Application number : 10-301992

(22)Date of filing : 23.10.1998

(71)Applicant : HITACHI LTD

(72)Inventor : MORIMOTO YASUTSUGU
MISHINA YUSUKE
KAJI HIROYUKI

(54) INFORMATION FILTERING SYSTEM AND PROFILE UPDATING METHOD IN THE SAME



(57)Abstract:

PROBLEM TO BE SOLVED: To make improvable the precision of its own profile by specifying one of other profiles storing the reference history information that has the higher degree of similarity to the reference history information than a prescribed level and updating its own profile based on the specified profile.

SOLUTION: A reference history file 5 stores the result of goodness of fit that is judged for the documents to which every user referred in the past as a reference history. An information filter program 32 compares the profile set for every user with the delivered documents and designates the documents having high degrees of coincidence with the set profile as the documents to be distributed. A degree of similarity judging program 34 compares the reference histories of users with each other and detects the users having the similar reference histories. A profile updating program 35 refers to the profiles of similar users extracted by the program 34 and updates the profile of each user.

Ref 4

対応なし、英抄

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2000－132559
（P2000－132559A）
(43)公開日　平成12年５月12日(2000.5.12)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード*（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/403	３４０Ａ	５Ｂ０７５
		15/40	３１０Ｆ	
		15/403	３５０Ｃ	

審査請求　未請求　請求項の数22　ＯＬ　（全 14 頁）

(21)出願番号　特願平10－301992

(22)出願日　平成10年10月23日(1998.10.23)

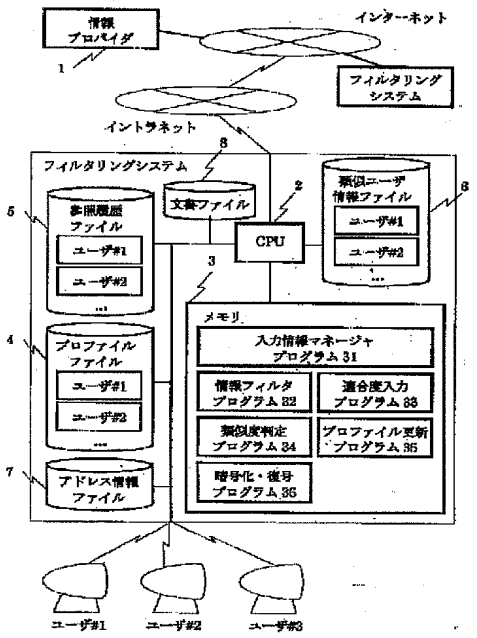
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　森本　康嗣
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(72)発明者　三科　雄介
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内
(74)代理人　100068504
弁理士　小川　勝男

最終頁に続く

(54)【発明の名称】　情報フィルタリングシステムにおけるプロファイル更新方法及び情報フィルタリングシステム

(57)【要約】
【課題】　情報フィルタリング用のプロファイルを精密化する。
【解決手段】　ユーザの興味を設定したプロファイルを用いて入力情報を適切なユーザに配信する情報フィルタリングシステムにおいて、類似した興味を持つユーザのプロファイルを用いることにより自分のプロファイルを精密化する。さらに、類似した興味を持つ他のユーザを自動的に決定することにより、効率的なプロファイル精密化を行う。

図1



【特許請求の範囲】
【請求項１】プロファイルに記述された情報要求と入力された電子化文書とを比較し、上記情報要求を満たす電子化文書を上記プロファイルのユーザに配信する情報フィルタリングシステムにおけるプロファイルの更新方法において、
上記プロファイルごとに上記配信された電子化文書の情報を参照履歴情報として保存し、
第一のプロファイルについて、その参照履歴情報が上記保存された他のプロファイルのうちその参照履歴情報との類似度が所定値よりも高い他のプロファイルを特定し、
上記特定された他のプロファイルに基づき上記第一のプロファイルを更新することを特徴とするプロファイル更新方法。
【請求項２】請求項１記載のプロファイル更新方法において、
上記参照履歴情報として、上記プロファイルに基づき配信された電子化文書の文書識別子と上記配信された電子化文書に対するユーザの興味との適合を示す適合度とを含み、
上記第一のプロファイルの参照履歴情報より得られる電子化文書のうち高い適合度が付与された電子化文書の集合と上記他のプロファイルの参照履歴より得られる電子化文書のうち高い適合度が付与された電子化文書の集合との重なりを上記類似度とすることを特徴とするプロファイル更新方法。
【請求項３】請求項２記載のプロファイル更新方法において、
上記適合度はユーザによる入力され、または配信された電子化文書のうちユーザにより全文が閲覧されたものを上記適合度が高いと擬制することを特徴とするプロファイル更新方法。
【請求項４】請求項２記載のプロファイル更新方法において、
上記第一のプロファイルのユーザに配信された電子化文書（Ａ）に対する適合度を、上記第一のプロファイルのユーザと類似した興味を有するユーザが上記電子化文書（Ａ）に付した適合度により擬制することを特徴とするプロファイル更新方法。
【請求項５】請求項１記載のプロファイル更新方法において、
上記他のプロファイルとして、所定のプロファイルを一つのプロファイルとしてにまとめたものを含むことを特徴とするプロファイル更新方法。
【請求項６】プロファイルに記述された情報要求と入力された電子化文書とを比較し、上記情報要求を満たす電子化文書を上記プロファイルのユーザに配信する情報フィルタリングシステムにおけるプロファイルの更新方法において、
第一のプロファイルの情報要求と他のプロファイルの情報要求との差分である差分タームを抽出し、
上記差分タームを上記第一のプロファイルにマージすることを特徴とするプロファイル更新方法。
【請求項７】請求項６記載のプロファイル更新方法において、
上記他のプロファイルは、上記第一のプロファイルのユーザにより指定される、もしくは上記第一のプロファイルの情報要求が上記情報フィルタリングシステムに登録されている他のプロファイルの情報要求との類似度が所定値よりも高いことにより特定されることを特徴とするプロファイル更新方法。
【請求項８】請求項６記載のプロファイル更新方法において、
上記抽出された差分タームは、上記第一のプロファイルにマージする差分タームの指定をユーザにより受けるため、表示画面上に表示されることを特徴とするプロファイル更新方法。
【請求項９】請求項８記載のプロファイル更新方法において、
上記他のプロファイルのユーザに配信されており、かつ上記第一のプロファイルのユーザには配信されていない電子化文書であって、上記差分タームを含むものを支援情報として抽出し、
上記抽出された支援情報は、ユーザによる差分タームの指定を支援するため、上記表示画面上に表示されることを特徴とするプロファイル更新方法。
【請求項１０】請求項９記載のプロファイル更新方法において、
上記情報フィルタリングシステムは、各プロファイルのユーザに関するユーザ情報をあらかじめ保持し、
上記他のプロファイルのユーザのユーザ情報を支援情報として抽出し、
上記抽出された支援情報は、ユーザによる差分タームの指定を支援するため、上記表示画面上に表示されることを特徴とするプロファイル更新方法。
【請求項１１】請求項１０記載のプロファイル更新方法において、
上記支援情報として抽出されるユーザ情報は、一部制限可能であることを特徴とするプロファイル更新方法。
【請求項１２】プロファイルに記述された情報要求と入力された電子化文書とを比較し、上記情報要求を満たす電子化文書を上記プロファイルのユーザに配信する情報フィルタリングシステムにおけるプロファイルの更新方法において、
第一のプロファイルの更新起動タイミングを決定し、
上記更新起動タイミングで、第一のプロファイルの情報要求と所定のプロファイルの情報要求との差分である差分タームを抽出し、
上記差分タームにより、上記第一のプロファイルを更新

することを特徴とするプロファイル更新方法。
【請求項１３】請求項１２記載のプロファイル更新方法において、
上記所定のプロファイルは、上記第一のユーザによって指定された他のユーザのプロファイル、または上記第一のプロファイルの情報要求が類似する他のユーザのプロファイルであることを特徴とするプロファイル更新方法。
【請求項１４】請求項１２記載のプロファイル更新方法において、
上記プロファイルごとに上記プロファイルに基づき配信された電子化文書の文書識別子と上記配信された電子化文書に対するユーザの興味との適合を示す適合度とを含む参照履歴情報を保存し、
上記所定のプロファイルは、上記第一のプロファイルとその参照履歴情報が類似するプロファイルであることを特徴とするプロファイル更新方法。
【請求項１５】請求項１２記載のプロファイル更新方法において、
上記各プロファイルの更新状況を監視し、上記所定のプロファイルの更新処理がなされたことが検出された時点で更新起動を決定することを特徴とするプロファイル更新方法。
【請求項１６】プロファイルに記述された情報要求と電子化文書とを比較し、上記プロファイルの情報要求を満たす電子化文書を判定する情報フィルタ機能と、
第一のプロファイルと第二のプロファイルとの類似度を判定するプロファイル類似判定機能と、
上記プロファイル類似判定機能により上記第二のプロファイルが上記第一のプロファイルに類似すると判定された場合、上記第二のプロファイルの情報要求により上記第一のプロファイルの情報要求を更新するプロファイル更新機能とを有することを特徴とするコンピュータ読取可能な記録媒体。
【請求項１７】請求項１７記載のコンピュータ読取可能な記録媒体において、
上記情報フィルタ機能により上記プロファイルの情報要求を満たすと判定された電子化文書に対して、ユーザの興味との適合の度合を示す適合度を入力する適合度入力機能を有し、
上記プロファイル類似判定機能において、上記適合度入力機能により入力された上記適合度により、上記第一のプロファイルと第二のプロファイルとの類似度を判定することを特徴とするコンピュータ読取可能な記録媒体。
【請求項１８】情報ネットワークに接続され、上記情報ネットワークを介して配送された電子化文書をフィルタリングする情報フィルタリングシステムにおいて、
複数のプロファイルを格納するプロファイル記録手段と、
上記プロファイルの情報要求を満たす上記電子化文書を判定する情報フィルタプログラムと、上記複数のプロファイル間の類似度を判定するプロファイル類似判定プログラムと、プロファイルの情報要求をその類似するプロファイルの情報要求により更新するプロファイル更新プログラムとを格納するメモリ手段と、
上記メモリ手段に格納されたプログラムを実行する処理手段と、
上記類似判定プログラムを上記処理手段により実行して得られた各プロファイルに類似するプロファイル情報を格納する類似プロファイル情報記録手段とを有することを特徴とする情報フィルタリングシステム。
【請求項１９】請求項１８の情報フィルタリングシステムにおいて、
上記メモリ手段は、上記プロファイルに基づき配信された電子化文書に対するユーザの興味との適合を示す適合度の入力を受ける適合度入力プログラムを格納しており、
上記プロファイルごとに上記プロファイルに基づき配信された電子化文書の文書識別子と上記適合度入力プログラムを上記処理手段により実行して得られた上記適合度とを含む参照履歴情報を格納する参照履歴記録手段とを有し、
上記参照履歴記録手段を参照し、第一のプロファイルにおいて高い適合度が付与された電子化文書の集合と第二のプロファイルにおいて高い適合度が付与された電子化文書の集合との重なりにより上記複数のプロファイル間の類似度を判定することを特徴とする情報フィルタリングシステム。
【請求項２０】情報プロバイダより情報ネットワークを介して電子化文書の配信を受け、上記電子化文書を各ユーザのプロファイルによりフィルタリングして、上記プロファイルに記述された情報要求を満たす電子化文書を配信する情報フィルタリング方法において、
所定のタイミングで第一のユーザに対して、そのプロファイルの更新起動を要請し、
上記要請に応じて、上記第一のユーザからのプロファイルの更新起動命令を受けた場合には、
上記プロファイルを第二のユーザのプロファイルに基づき更新することを特徴とする情報フィルタリング方法。
【請求項２１】請求項２０記載の情報フィルタリング方法において、
上記第二のユーザは、上記第一のユーザによって指定された他のユーザ、または上記プロファイルの情報要求が類似するプロファイルを所有する他のユーザであることを特徴とする情報フィルタリング方法。
【請求項２２】請求項２０記載の情報フィルタリング方法において、
上記プロファイルごとに上記プロファイルに基づき配信された電子化文書の文書識別子と上記配信された電子化文書に対するユーザの興味との適合を示す適合度とを含

む参照履歴情報を保存し、
上記第二のユーザは、上記第一のユーザのプロファイルとその参照履歴情報が類似するプロファイルを有する他のユーザであることを特徴とする情報フィルタリング方法。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、情報フィルタリングシステムに関する。特に、情報フィルタリングに用いられるユーザのプロファイルの精密化方法に関する。
【０００２】
【従来の技術】現在、ネットワークの発達、特にインターネットの普及に伴って、ネットワークを介して入手可能な電子化文書の量が膨大になってきている。インターネット上の電子化文書に対するアクセス方法は、従来、ｈｔｔｐ(hyper text transferprotocol)によるWWW(World Wide Web)文書のブラウジングが中心であった。WWW文書のブラウジングは、Ｗｅｂブラウザと呼ばれるソフトウェア上で、そのWWW文書のアドレスであるＵＲＬ(Uniform Resource Locator)を指定することによって行う。ＵＲＬは、そのWWW文書が存在する計算機及びその計算機上でのディレクトリなどを一意に指定する。ユーザが個別にWWW文書を指定してアクセスするブラウジングは、通常、プル型と呼ばれている。
【０００３】これに対して、プッシュ型と呼ばれる情報配信型のサービス（ユーザは個別にWWW文書を指定することなくアクセスする）が注目されている。このような情報配信型のサービスでは、各ユーザが興味を持つ情報（WWW文書）を配信することが重要である。
【０００４】情報フィルタリングシステムは、各ユーザが興味をもつ事柄についての情報を記述したデータ（インタレストプロファイル(以下、「プロファイル」)と呼ばれる）に基づいて、配信される情報をフィルタリングするシステムである。情報フィルタリングシステムは、WWW文書がネットワークを通じて配信されると、配信されたWWW文書と各ユーザのプロファイルとを比較することにより、そのWWW文書に興味を持つであろうユーザを決定し、そのユーザに対してのみにそのWWW文書を配送する。これにより、ユーザは、自分が興味を持つWWW文書のみを受け取ることができる。
【０００５】したがって、情報フィルタリングにおいて最も大きな課題は、配信されたWWW文書がユーザの興味に合致するかどうかを正しく判定することであり、そのためには、プロファイルにユーザの興味が正確に記述されている必要がある。プロファイルを精密化するために、特開平９－１５３０６４号公報に開示されているレリバンスフィードバック技術が提案されている。レリバンスフィードバックでは、プロファイルに基づいて各ユーザに配信された情報に対して、ユーザが適合度（ユーザの興味に対して配信された情報が合致する程度）を入力する。システムは、入力された適合度をプロファイルにフィードバック（例えばプロファイルに付された重みを変更する）することによってプロファイルがユーザの興味をよりよく反映するように精密化する。
【０００６】また、別のアプローチとして特開平９－２６５４７８号公報に開示されている協調フィルタリング技術が提案されている。このアプローチは自分と同じ興味を有する他のユーザを発見し、他のユーザに配信された情報は自分にも同様に配信してもらうことにより、興味のあるWWW文書をもれを少なくして受け取るようにするものである。
【０００７】
【発明が解決しようとする課題】レリバンスフィードバック技術では、プロファイルの精密化をユーザ毎に行う。しかし、各ユーザを個別に扱っている限り、精密化の精度には限界があった。従来のレリバンスフィードバックでは、各ユーザが参照した文書にもとづいて自己のプロファイルの精密化を行う。しかし、ユーザ各人が参照できる文書数には限りがあるため、プロファイル精密化の精度には限界がある。
【０００８】また、協調フィルタリング技術では、自己のプロファイル自体は修正されず、また、他人の興味が自分の興味と完全に一致していることはまれであるから配信されるWWW文書量の増大が避けられない。
【０００９】本発明の目的は、以上の従来技術の問題を踏まえて、新しいプロファイル精密化方法およびそれを適用した情報フィルタリングシステムを提供する。本発明では、同じ情報フィルタリングシステムを利用する他のユーザの知識を利用して自己のプロファイルを精密化する。
【００１０】
【課題を解決するための手段】このため、プロファイルに記述された情報要求と入力された電子化文書とを比較し、情報要求を満たす電子化文書をプロファイルのユーザに配信する情報フィルタリングシステムにおいて、プロファイルごとに配信された電子化文書の情報を参照履歴情報として保存し、あるプロファイルについて、その参照履歴情報が保存された他のプロファイルのうちその参照履歴情報との類似度が所定値よりも高い他のプロファイルを特定し、特定された他のプロファイルに基づきプロファイルを更新する。参照履歴情報は、プロファイルに基づき配信された電子化文書の文書識別子と配信された電子化文書に対するユーザの興味との適合を示す適合度とを含むようにする。
【００１１】
【発明の実施の形態】以下，本発明の一実施例を図面を用いて説明する。
【００１２】図１に、本発明の情報フィルタリングシステムの実施例の全体構成を示す。情報フィルタリングシステムは、インターネットまたはイントラネットに接続

されている。システムは、インター／イントラネットを介して情報プロバイダ１から送信された情報（WWW文書等）をフィルタリングし、各ユーザに配送する。情報プロバイダ１は、情報を送信するサービスの提供者であり、ネットワーク上には複数のプロバイダが存在しうる。情報フィルタリングシステムには、異なるリソースからの情報、例えば、複数のプロバイダから送信された情報やインター／イントラネットなどからInternet Robotと呼ばれるようなエージェント（代理人）システムを用いて収集された情報が配信される。

【００１３】ＣＰＵ２は、メモリや各種ファイル上にあるプログラムやデータを用いて情報フィルタリング処理を行う中央処理装置である。メモリ３には、各種の処理を行うプログラムやデータがロードされる。プログラムは、図示しない例えば磁気記録媒体、光記録媒体のようなコンピュータ読取可能な記録媒体に格納されてシステムに供給され、メモリ３にロードされる。また、システムは、プロファイルファイル４、参照履歴ファイル５、類似ユーザ情報ファイル６、アドレス情報ファイル７、文書ファイル８を有する。

【００１４】プロファイルファイル４には、ユーザごとのプロファイルが格納されている。プロファイルの一例を図２に示す。この例では、プロファイルとして単語を設定する。図２の例に示されるプロファイルを持つユーザは、暗号関係の情報に興味があり、暗号に関連する語として、「暗号」、「公開鍵」、「ＰＧＰ」、「セキュリティ」という語を設定している。

【００１５】参照履歴ファイル５は、ユーザごとに、ユーザが過去に参照した文書について適合度を判定した結果を、参照履歴として格納する。各ユーザの参照履歴の一例を図３に示す。参照履歴は、各ユーザに配送された文書ＩＤとその適合度との組の集合である。ユーザ番号は、情報フィルタリングシステムのユーザのＩＤ番号であり、各ユーザに対して配信された文書ＩＤが参照文書ＩＤとして登録される。適合度は、ユーザが入力した適合度を数値化したものであり、「１」が適合している（ユーザの興味と合致している）場合を、「０」が適合していない（ユーザの興味と合致していない）場合を示す。

【００１６】類似ユーザ情報ファイル６は、各ユーザごとに、興味の類似するユーザ（「類似ユーザ」という）の情報を格納する。各ユーザの参照履歴が類似しているユーザはその興味が類似するとみなして、そのＩＤを類似ユーザとして登録する。図４の例では、ユーザ＃１の類似ユーザとして、ユーザ＃２、＃７が登録されている。

【００１７】アドレス情報ファイル７は、協調して動作する他のフィルタリングシステムの所在を例えばIP-addressなどの形で格納する。例えば、大きな企業などでは、物理的に離れた個所に事業所／営業所などが存在する場合がある。このような場合に、フィルタリングシステムをそれぞれの事業所（営業所）に設け、各フィルタリングシステム同士で所有するデータのやり取りが行えれば便利である。そのため、対象となるシステムの所在をアドレス情報ファイルに記述しておく。

【００１８】文書ファイル８は、各ユーザごとに情報配信用に各ユーザのみが参照可能なファイルを有する。ユーザは情報フィルタリングシステムに接続された端末から、それぞれのファイルにアクセスすることにより、フィルタリングされた文書の配信を受ける。

【００１９】メモリ３に格納されるプログラムについて説明する。入力情報マネージャプログラム３１は、情報プロバイダ、インター／イントラネットから得られる情報を管理する。例えば、複数のリソースから同時に多くの情報を受信した場合などに、計算機に対する負荷が大きくなり過ぎるのを防止するため、図示しないファイルに情報を一旦バッファリングすることにより、情報フィルタリングシステムの負荷を調整する。また、様々なリソースから得られた情報を一括して管理し、配信された文書に対して文書ＩＤを付与する。

【００２０】情報フィルタプログラム３２は、各ユーザごとに設定されたプロファイルと配送された文書との比較を行い、プロファイルに合致する程度が高い文書を配信すべき文書として指定する。

【００２１】適合度入力プログラム３３は、配信された文書に対し、各ユーザにより適合度の入力を受ける。適合度は、配信された文書が各ユーザの要求と合致している程度を表わす。適合度は、ユーザの要求に合致するか否かの二者択一で入力する。５段階評価などで入力することも可能である。以下の例では、yes/noの二者択一で入力する（図６参照）。

【００２２】類似度判定プログラム３４は、各ユーザの参照履歴を比較し、類似する参照履歴を持つユーザを発見する。

【００２３】プロファイル更新プログラム３５は、類似度判定プログラム３４によって抽出された類似ユーザのプロファイルを参照して、各ユーザのプロファイルを更新する。

【００２４】暗号化・復号プログラム３６は、情報をネットワークを介して送信する際、情報の暗号化／復号化を行う。

【００２５】以下、本実施例での情報フィルタリングシステムにおける各処理を説明する。

【００２６】（１）文書の配信・表示処理（図５）
情報フィルタリングシステムは、文書が配送されているかどうかをチェックする（ステップ１１）。情報フィルタプログラム３２を起動し、プロファイルファイル４に格納された各ユーザのプロファイルと配送された文書とを比較し、配信するユーザを決定する（ステップ１２）。ステップ１２は次のような手順で行われる。ま

ず、配送された文書を形態素解析し、タームに分割し、助詞などの機能語を削除して配送文書に含まれるターム集合ＩＴを作成する。次に、プロファイルファイル４から、各ユーザのプロファイルを逐次取り出し、各プロファイル中に格納されているターム集合ＰＴ（以下、「プロファイルターム」）とＩＴを比較することによって、プロファイルと配送された文書との一致度Ｍを計算する。

【００２７】Ｍ＝¦ＰＴ∩ＩＴ¦／¦ＰＴＵＩＴ¦

Ｍの値が予め定められた閾値より大きいユーザを配信の対象とする。

【００２８】なお、一致度Ｍの計算式は上記の数式に限定されない。また、比較対照とするタームは、語彙の特殊性または出現の頻度などによって限定することも可能である。

【００２９】配信先と判定されたユーザに対し文書を配信する（ステップ１３）。本実施例では、文書ファイル８中の配信先ユーザのファイルにその文書を格納する。情報の配信方法としては、任意の方式が採用できる。

【００３０】各ユーザから、配送された文書を表示する要求があったかどうか調べる（ステップ１４）。表示要求がなければ、ステップ１１に戻って待機する。表示要求を受けると、文書ファイル８中の配信文書の表示を要求したユーザのファイルから、配信された文書を取り出し、端末に表示する処理を行う（ステップ１５）。ユーザに表示される端末画面の例を図６に示す。文書ファイル８に格納された配信文書には、最新（もしくは未読）を示すフラグをたてておき、表示要求を受けると最新の(もしくは、未読の)情報を表示する。各ユーザは配信文書に適合度を入力する。（ステップ１７）。例えば、図６において、記事ごとに配置された「適合」ボタンによって適合度を入力する。配信文書がユーザの興味に適合していれば「適合」ボタンを押下し、そうでなければ何もしない。このように入力された適合度は数値化され、配信文書ＩＤとともにメモリの所定のテーブル（図示せず）に一時的に保存される（ステップ１８）。配信文書が複数存在する場合には、この処理を繰り返す。ユーザが配信文書の表示を終了すると、一時保存されている配信文書ＩＤと適合度とを参照履歴ファイル５に格納する（ステップ１９）。

【００３１】本実施例では、参照履歴はプロファイルに対応して格納されている。プロファイルはユーザの特定の興味に対応して設定され、その興味は経時的に変化しないものであるという前提に立つ。プロファイルはその特定の興味をよりよく反映するように洗練されていく。したがって、参照履歴から類似ユーザを発見するためにはデータ量が多いほど望ましいところから、参照履歴はプロファイルの精密化処理の実行にかかわらず、継続的に蓄積されていくことが望ましい。しかし、参照履歴を格納しているハードディスクなどの記憶装置の容量には限りがあるため、ファイルにデータを格納しようとした際にあふれ（容量不足）が発生する場合がある。以下では、このような場合に対処するための代表的な方法について図７の処理フローを用いて説明する。

【００３２】参照履歴を格納する際、ディスクのあふれがないかどうか調べる（ステップ１９１）。あふれがなければ、参照履歴を格納する（ステップ１９２）。あふれが検出された場合には、圧縮処理を行う。圧縮処理の対象となるユーザを決定する（ステップ１９３）。圧縮処理対象とするユーザは、大きな参照履歴を持つユーザまたは使用頻度が低いユーザを対象とできる。圧縮処理対象とするユーザの参照履歴を圧縮する（ステップ１９４）。数回圧縮を繰り返しても、あふれが発生する場合には処理を停止し、管理者に連絡するなどの対策を行う。

【００３３】参照履歴の圧縮は、参照履歴中の優先度が低い文書ＩＤの情報を削除する。各文書ＩＤの優先度は、日時（新しいものほど優先度を高くする）によって判定する。あるいは、その文書がどの程度、情報フィルタリングシステムのユーザに配信されたか（配信されたユーザの多いものほど優先度を高くする）によって判定する。これは、参照履歴が類似ユーザの判定に用いられるものであり、その文書を参照した他のユーザが少ないければその判定にほとんど影響しないためである。

【００３４】なお、参照履歴ファイル５の全体でなく、各ユーザごとの参照履歴ファイルに上限を設けておき、上限に達した時点で、参照履歴ファイル圧縮を行うようにしてもよい。この方法では、参照履歴ファイル全体の大きさの上限が定められる。

【００３５】（２）プロファイルの精密化処理（図８）

プロファイル精密化処理を起動する（ステップ２１）。タイマーなどにより定期的に起動してもよく、前回の精密化処理後に蓄積された参照履歴が一定以上の大きさになったことを検知して起動してもよい。また、類似ユーザ情報ファイル６を参照し、類似ユーザのプロファイルについて変更があった場合に、処理を起動することもできる。

【００３６】プロファイルの精密化はあらかじめ設定されているため、類似ユーザを探索する処理はフィルタリング処理とは独立して行うことができる。その意味で、一般の情報検索と異なり、検索条件の精密化にはリアルタイム性を要しない。そのため、類似ユーザの探索処理には計算負荷が大きいが、実用上の問題はない。また、ユーザによる精密化起動指示を待たず、上述のような条件により処理起動もしくはメール等により起動要請を通知することが望ましい。これは、情報フィルタリングの文書ファイル８の容量の問題から、フィルタリング処理は文書が配送される度に実行せざるを得ない。この場合、長期にわたってプロファイルの更新を実行されないことにより、ユーザの興味に合致する文書が配送されな

くなるおそれがあるためである。

【００３７】精密化処理を要求する要求ユーザの参照履歴Ｒｂを参照履歴ファイル５から取得する（ステップ２２）。また、要求ユーザのプロファイルをプロファイルファイル４から取得する（ステップ２３）。

【００３８】その後、他のユーザ（「比較対照ユーザ」という）の参照履歴Ｒｔを参照履歴ファイル５から取得する（ステップ２５）。ここで、比較対照ユーザは、自情報フィルタリングシステムのみではなく、アドレス情報ファイル７に記述された他の情報フィルタリングシステムをも対象とすることが望ましい。なお、異なるシステム間で情報をやり取りする際には、送信側の計算機の暗号化・復号プログラム３６によって暗号化を行った後で送信し、受信側の計算機の暗号化・復号プログラム３６によって復号する。

【００３９】要求ユーザの参照履歴Ｒｂと比較対象ユーザの参照履歴Ｒｔとを比較し、類似度を決定する（ステップ２６）。比較は、次のように行う。参照履歴ＲｂおよびＲｔから、適合度が「１」の文書ＩＤ集合である適合文書履歴Ｒｂ′、Ｒｔ′を抽出する。このとき、類似度ｓｉｍを次のように定める。

【００４０】

$$sim = |Rb' \cap Rt'| / |Rb' \cup Rt'|$$

類似度では、同一文書に関する適合度により類似度を定義しているが、情報フィルタリングシステムに入力される情報の量が少ない場合には、類似度をうまく設定できない場合がある。別の類似度を求めるアプローチとして、参照履歴中の文書間の類似性を求め、類似文書に関する適合度を用いることで類似度を定義することも可能である。類似文書の判定は、例えば、文書中の単語の出現頻度を用いて各文書を特徴付ける単語の集合(特徴語集合)を抽出し、特徴語集合同士を比較することによって行える。２個の文書Ｄ１、Ｄ２から特徴語集合Ｆ１、Ｆ２を抽出する。特徴語集合は、各文書から形態素解析処理によって単語を抽出し、名詞、動詞などの内容語の出現頻度をカウントし、予め定められた閾値以上の頻度を持つ内容語を抽出することによって得られる。特徴語集合の文書類似度Ｆｓｉｍは例えば、以下の式を用いる。

【００４１】

$$Fsim = |F1 \cap F2| / |F1 \cup F2|$$

文書類似度Ｆｓｉｍが予め定められた閾値以上の場合、文書Ｄ１とＤ２とは類似していると判定する。そして、類似度ｓｉｍの式における$|Rb' \cap Rt'|$の代わりに、類似する文書数によって類似度ｓｉｍを定義する。また、類似文書の判定には、他にも、ｔｆ－ｉｄｆ(term frequency - inverse document frequency)法を用いたベクトル空間法などを用いることもできる。他の例については、「J. Mostafa et al., A multilevel Approach to Intelligent InformationFiltering: Model, System, and Evaluation, ACM Transactions on Information Systems, Vol. 15, No.4, pp.368-399, October 1997.」などに開示されているため詳細な説明は省略する。

【００４２】参照履歴の類似度ｓｉｍが予め定められた閾値と比較し（ステップ２７）、類似度ｓｉｍの大きい比較対象ユーザは要求ユーザと類似した興味を持っていると判定し、対応するプロファイルをプロファイルファイル４から取得し、比較対象ユーザのプロファイルから要求ユーザのプロファイルに含まれないタームを抽出して、追加タームリスト（図示せず）に格納する（ステップ２８）。全ての他のユーザについて以上の処理が終了すれば（ステップ２４）、得られた追加タームリスト中のタームを要求ユーザのプロファイルに追加し、新規プロファイルとしてプロファイルファイルに格納する（ステップ２９）。

【００４３】なお、プロファイルとしてタームの論理式（検索式）が与えられている場合には次のようにする。第一の方法としては、プロファイルの論理式からタームのみを切り出し、タームの集合について同じ処理を行う。追加されるタームは、今までの検索式に、ＯＲ（論理和）で追加する。第二の方法としては、ステップ２８で、追加する検索式として各ユーザの検索式を全て抽出し、ステップ２９で、類似ユーザのプロファイルの検索式をＯＲで追加する。この場合、例えば、自分のプロファイルが「暗号AND認証」であり、類似ユーザのプロファイルが「暗号ANDセキュリティ」であれば、新しいプロファイルは、「（暗号AND認証）OR（暗号ANDセキュリティ）」となる。得られた検索式を論理式の簡略化方式に従って簡略化してもよい。

【００４４】プロファイルを精密化にあたっては、追加タームリストをユーザに表示し、実際にプロファイルに追加するかどうかをユーザに判断させるような構成とすることも可能である。その場合、ユーザの判断を支援するための情報を提示することが望ましい。以下では、ステップ２８におけるインタラクティブな追加ターム決定方法について図９を用いて説明する。

【００４５】追加タームリスト中のタームを表示する（ステップ２８１）。表示画面の例を図１０に示す。図１０では、追加タームとして、「認証」、「電子署名」、「楕円関数」が表示されている。ユーザは、これらの追加タームを参照し、それぞれのタームについて自分のプロファイルに追加するかどうかを「yes」／「no」で判断し、追加ボタンによってシステムに追加の可否を指示する。

【００４６】ユーザは、判断の参考になる情報を望む場合は、支援情報ボタンを押下する。支援情報ボタンが押下された場合には、追加ターム判定のための支援情報を表示する（ステップ２８４）。図１１に、追加ターム判定支援情報の表示例を示す。支援情報は、プロファイルに追加タームを加えることによって、新たにどのような

文書が入手できるかを例示するものである。「認証」という追加タームに関し、「認証」を含み、類似ユーザのみに配信された文書＃３００８が表示されている。ユーザは、この文書を読み、「認証」という追加タームを追加するかどうかの適切さを判断する。この処理は、要求ユーザに配信された文書ＩＤリストと類似ユーザに配信された文書ＩＤリストの差分を抽出し、追加タームのそれぞれについて抽出された文書を探索し、各タームごとに各タームを含む文書ＩＤリストを表示することにより実現できる。

【００４７】また、他の支援情報として、類似ユーザのユーザ情報を表示してもよい。ステップ２７（図８）において類似度が高いと判定されたユーザのユーザ情報を取得する。ユーザ情報とは、例えば、ユーザの氏名、所属部署、業務内容、電子メールアドレス、電話番号、住所などである。そして、このユーザ情報を追加タームリストと共に表示する。この場合、各ユーザは、支援情報として表示可能なユーザ情報に制限を加える。例えば、社内などで秘密のプロジェクトなどに参加しているユーザなどの場合である。また、インターネット上で広く利用されるシステムなどにおいても、プライバシー上の問題がある。そこで、各ユーザは他のユーザに対して開示してもよい情報の範囲を予め定めておく。一切フィルタリング情報を開示しない場合は、プロファイルに対して、プロファイル精密化処理の対象としても良いかどうかを示すフラグを立てておき、このフラグが立てられているプロファイルに対応する参照履歴は、ステップ２５の参照履歴取得処理の対象外とする。他にも、プロファイルを参照することは可能だが、ユーザの氏名などは秘密にするなどのバリエーションが考えられる。

【００４８】追加ボタンが押下されれば（ステップ２８２）、ユーザが選択した追加タームが残るように追加タームリストを変更する（ステップ２８５）。

【００４９】なお、ステップ２６（図８）の参照履歴間の類似度ｓｉｍの判定処理において、比較対照ユーザの参照履歴として適合度が「１」の文書のみ（適合文書履歴Ｒｔ′）を対象とした。これに対して、配信された全ての文書を含むＲｔを用いることも可能である。比較対象ユーザの参照履歴として、ＲｔとＲｔ′を用いる場合では、以下のような違いがある。

【００５０】Ｒｔは、比較対照ユーザが設定しているプロファイルによって配信される情報全てであり、Ｒｔとプロファイルとは一対一に対応している。 よって、Ｒｂ′（要求ユーザの適合文書履歴）とＲｔが類似していれば、Ｒｔと対応するプロファイル中のタームは、要求ユーザにとって追加すべきタームであるといえる。しかし、Ｒｔ中に比較対象ユーザにとっても興味に合わない文書が含まれている可能性があるため、これらのノイズが類似ユーザを探す場合に精度を下げる可能性がある。

【００５１】一方、Ｒｔ′は、比較対象ユーザが興味を持つ文書のみが含まれている。よって、Ｒｂ′とＲｔ′とが類似していれば、比較対象ユーザの興味と要求ユーザの興味とは一致している可能性が非常に高い。しかし、比較対照ユーザが興味をもたなかった文書に対応するタームが追加される可能性が生じることで精度を下げる可能性がある。そのため、いずれの参照履歴（ＲｔまたはＲｔ′）を用いるかは、目的に応じて決定すればよい。

【００５２】配信文書に対するユーザによる適合度の入力の手間を軽減する例を説明する。

【００５３】第一の方法としては、ステップ２６（図８）の類似度判定処理において、配信された文書の文書ＩＤにより、ユーザ間の類似度を決定することが可能である。すなわち、ｓｉｍ＝┆Ｒｂ∩ Ｒｔ┆／┆ Ｒｂ∪ Ｒｔ┆として算出する。この場合、要求ユーザの適合文書履歴を用いる場合と比較すると精度が若干低下する可能性があるが、追加タームをチェックすることにより、適切なタームをプロファイルに追加するようにする。

【００５４】第二の方法としては、配信文書の表示と適合度の入力とを連動させることにより、明示的な適合度の入力をなくす。この場合の表示方法の例を図１２に示す。図１２は、配信文書のタイトル一覧を示している。タイトルがなければ先頭から予め定められた文字数のみ表示する。あるいは、「田中他、朝倉日本語新講座運用２・人文系研究のための言語データ処理入門、朝倉書店(1983)」などに開示されているような、重要語や重要文を抽出する技術によって抽出される重要語や重要文を記事と対応して表示しても良い。ユーザは、ステップ１５（図５）において、記事の一覧を参照して、自分が読みたいと判断した文書を選択し、選択した記事に対応する「表示」ボタンを押下することにより、記事の内容を表示させることができる。ここで、ユーザが選択した文書は適合と、選択されなかった文書は不適合とみなす。

【００５５】第三の方法としては、類似ユーザの適合度情報を利用し、ユーザが適合／不適合の判定をしなかった文書に対しても、類似ユーザの判定結果を援用して適合度情報を与えることも可能である。ユーザＡの類似ユーザとしてユーザＢが類似ユーザ情報ファイル６に登録されているとする。ユーザＡが別途、ユーザＢを指定しても良い。配信文書中、ユーザＡが適合度を指定しなかった文書がユーザＢの参照履歴中に存在するかどうかを調べる。存在した場合には、ユーザＢの参照履歴中の適合度を調べ、ユーザＢによる適合度をユーザＡの適合度として援用し、参照履歴に格納する。

【００５６】なお、各ユーザのプロファイルは別々に扱うのではなく、類似する興味を持つユーザを一つのグループとして扱うことも可能である。類似する興味をもつユーザは、以上で説明した実施例に従って自動抽出しても良いし、人手で設定しても良い。例えば、同じ業務を担当する複数のユーザは、最初から類似ユーザとして指

定することが可能である。既に類似した興味を持つユーザが分かっている場合には、類似する興味を持つユーザのプロファイルおよび参照履歴をマージしたデータを一時的に作成し、以上で説明した実施例を適用することにより、さらに他の類似した興味を持つユーザを精度良く発見することができ、プロファイルを精密化することが可能となる。

【００５７】さらに、本実施例では、説明を簡単にするため、各ユーザのプロファイルは１個であるように説明したが、実際にはユーザが複数の事柄に興味を持つことは多い。このような場合には、参照履歴をユーザ単位ではなく、プロファイル単位に作成することで、以上で説明した方法が適用可能である。図１３に複数のプロファイルを設定した場合の参照履歴の例を示す。ユーザ＃１が２個のプロファイルを持っており、各プロファイルに対して、参照履歴が設定されている。

【００５８】（３）プロファイルの精密化処理（図１４）

次に、プロファイルの精密化処理の別の実施例として、参照履歴を使用しないプロファイル精密化方法について説明する。本実施例では、プロファイル同士を比較することによってユーザの興味が類似しているかどうかを判定する。この方法の場合、参照履歴を使用しないため、処理が単純であり、ディスクの容量が圧縮できるなどの利点がある。図８に示した第一の例と共通する部分については説明を省略する。

【００５９】要求ユーザのプロファイルＰｂをプロファイルファイル４から取得する（ステップ３２）。比較対象ユーザのプロファイルＰｔをプロファイルファイル４から取得する（ステップ３４）。要求ユーザのプロファイルＰｂと比較対象ユーザのプロファイルＰｔを比較し、類似度ｓｉｍｐを決定する（ステップ２６）。

【００６０】ｓｉｍｐ＝|Ｐｂ∩Ｐｔ|／|Ｐｂ∪Ｐｔ|プロファイルの類似度ｓｉｍｐが予め定められた閾値より大きいかどうかを判定し（ステップ３６）、大きい場合には追加タームリストの設定を行う（ステップ３７）。

【００６１】プロファイル同士を直接比較する場合、設定されているターム数が少ないため、類似度を正しく求めることができない場合がある。これを解決するため、シソーラスを用いる。シソーラスとは、語と語の意味的関係を記述したデータである。関係としては、上位概念・下位概念、類義などがある。シソーラス上で所定の関係を持つ語（例えば、類義語等）を同じ語とみなすことにより、類似度を計算することができる。

【００６２】さらに、ユーザが自分と類似した興味を持つユーザをあらかじめ認識していれば、そのような類似ユーザをユーザが直接指定して、自己のプロファイルを精密化するようにすればよい。例えば、企業においては、自分と類似または関連する業務を行っている部署や人からそのような類似ユーザを特定できる。あるいは、実施例で述べたようなプロファイル精密化処理によって、あるユーザが自分と類似した興味を持っていることが分かると同時に、その類似ユーザの所属する部署の業務が自分の興味と関連が深いことが分かったとする。このような場合には、その部署に所属する他のユーザのプロファイルも参考にして自分のプロファイルを精密化できる。この場合には、図８または図１４の処理フローにおいて、あらかじめ比較対象ユーザとして指定する単数又は複数のユーザもしくはユーザが属するグループの名称を指定するステップを設ける。そして、プロファイルの精密化に利用する比較対象ユーザをあらかじめ指定されたユーザもしくはグループに属するユーザに限定してステップ２４～２９（図８）、ステップ３３～３９（図１４）の処理を行う。

【００６３】

【発明の効果】本発明によれば、情報フィルタリングシステムにおいて、自分と類似した興味を持つユーザが設定したプロファイルを利用することにより、自分のプロファイルをより自分の興味に適応したものに修正することが可能となる。

【図面の簡単な説明】

【図１】本発明の実施例である情報フィルタリングシステムのブロック図である。

【図２】プロファイルの例である。

【図３】参照履歴の例である。

【図４】類似ユーザ情報の例である。

【図５】情報配信・表示処理の処理フローである。

【図６】配信情報の表示例である。

【図７】記憶装置のあふれ対策処理の処理フローである。

【図８】プロファイル精密化処理の処理フローである。

【図９】インタラクティブな追加ターム決定処理の処理フローである。

【図１０】追加タームリストの表示例である。

【図１１】追加ターム判定支援情報の表示例である。

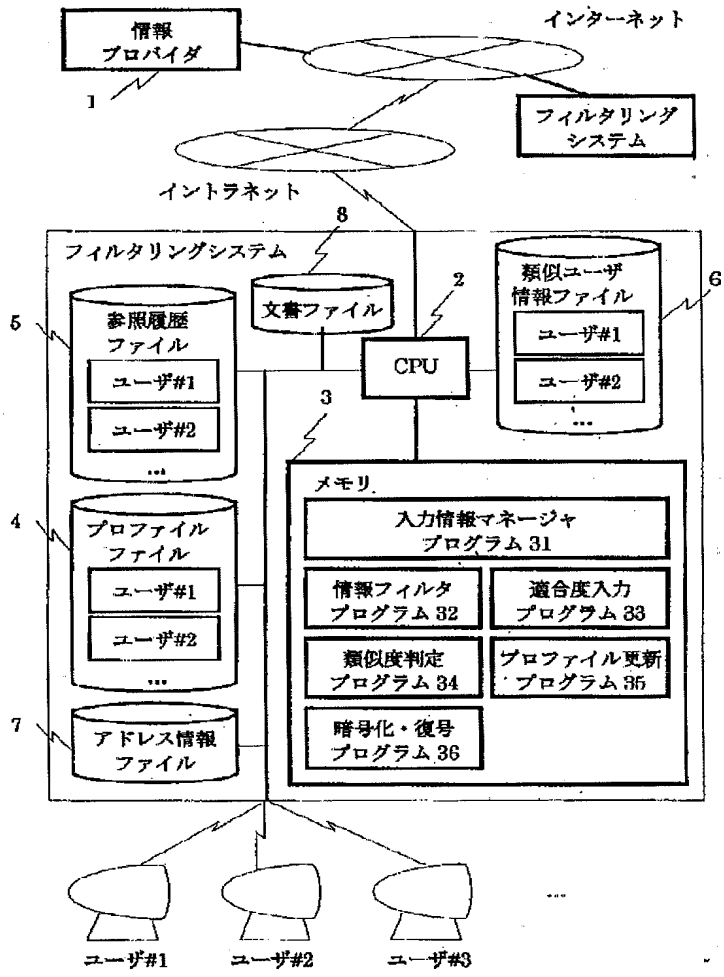
【図１２】明示的な適合度入力を省略するための配信情報の一覧表示画面の例である。

【図１３】各ユーザが複数のプロファイルを設定可能な場合の参照履歴の例である。

【図１４】第２の実施例におけるプロファイル精密化処理の処理フローである。

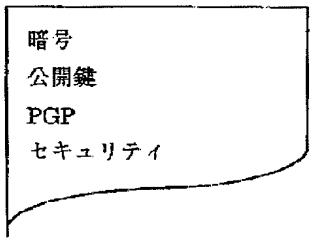
【図1】

図1



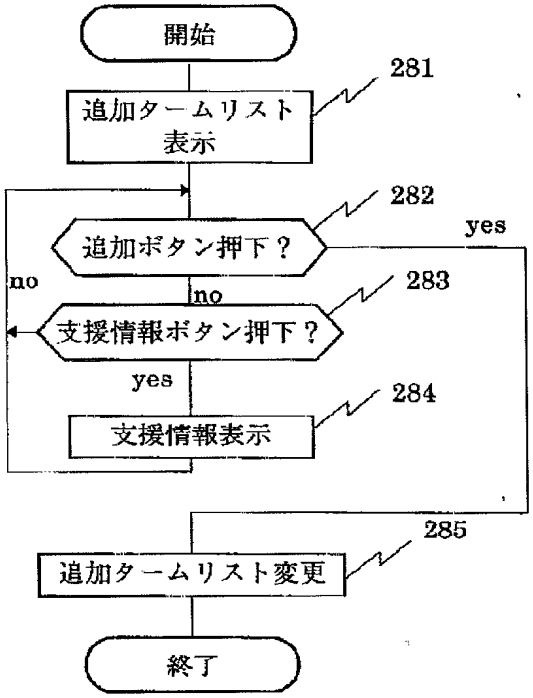
【図2】

図2



【図9】

図9



【図3】

図3

ユーザ番号					
#1	参照文書ID	#1024	#1248	#2045	#3042
	適合度	1	0	1	1
#2	参照文書ID	#1024	#2045	#3042	#3045
	適合度	1	1	1	0
#3	参照文書ID	...	...	...	...
	適合度	...	...	...	...

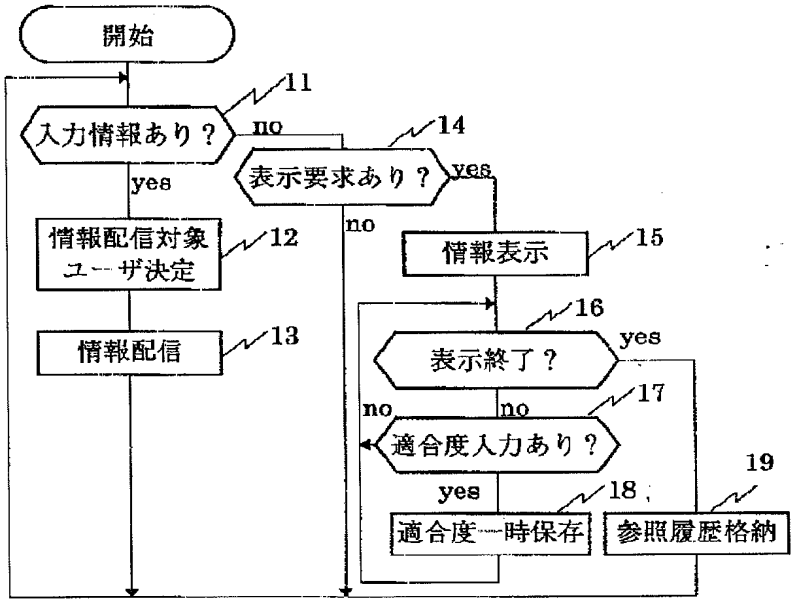
【図4】

図4

比較元ユーザ	類似ユーザ		
#1	#2	#7	—
#2	#4	—	
#3	—		
…	…	…	…

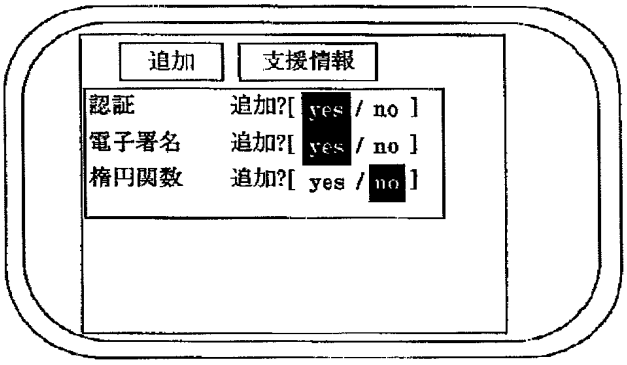
【図5】

図5



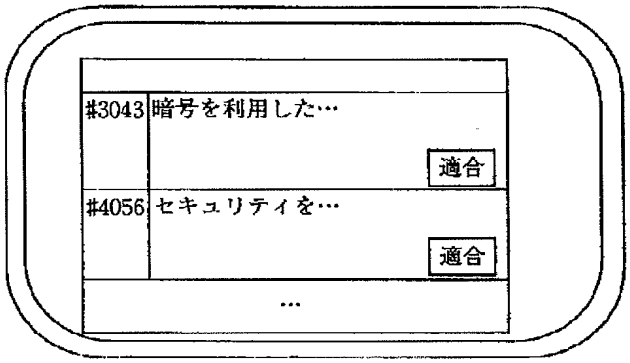
【図10】

図10



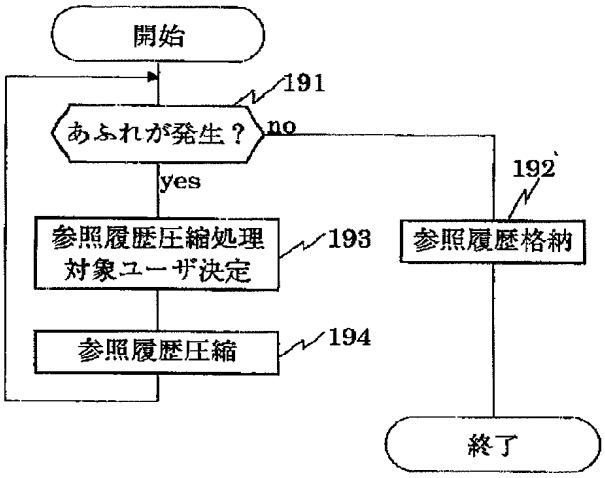
【図６】

図６



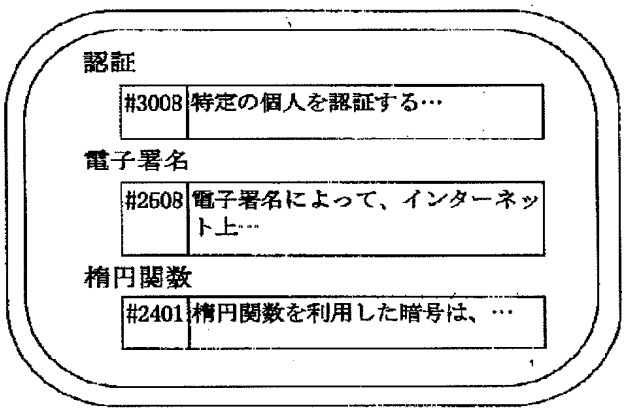
【図７】

図７



【図１１】

図１１



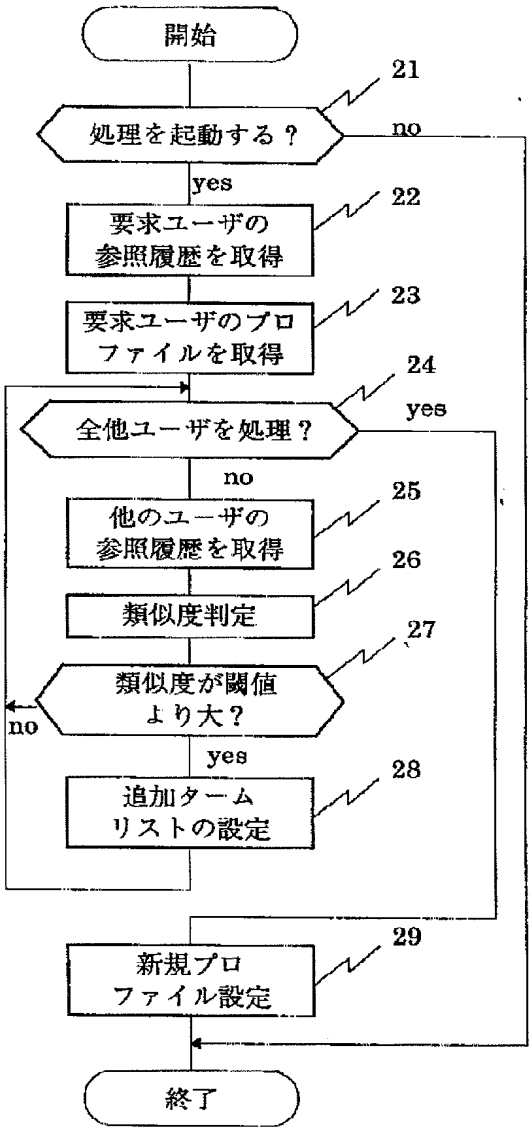
【図１３】

図１３

ユーザ番号	ﾌﾟﾛﾌｧｲﾙ番号					
#1	#1	参照文書 ID	#1024	#1248	#2045	#3042
		適合度	1	0	1	1
	#2	参照文書 ID	#1023	#2022	#3022	#3045
		適合度	1	0	1	0
#2	#1	参照文書 ID	…	…	…	…
		適合度	…	…	…	…

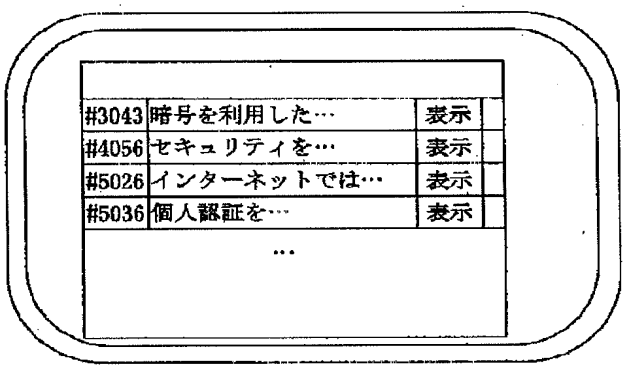
【図８】

図８


開始
21
処理を起動する？
no
yes
22
要求ユーザの
参照履歴を取得
23
要求ユーザのプロ
ファイルを取得
24
全他ユーザを処理？
yes
no
25
他のユーザの
参照履歴を取得
26
類似度判定
27
類似度が閾値
より大？
no
yes
28
追加ターム
リストの設定
29
新規プロ
ファイル設定
終了

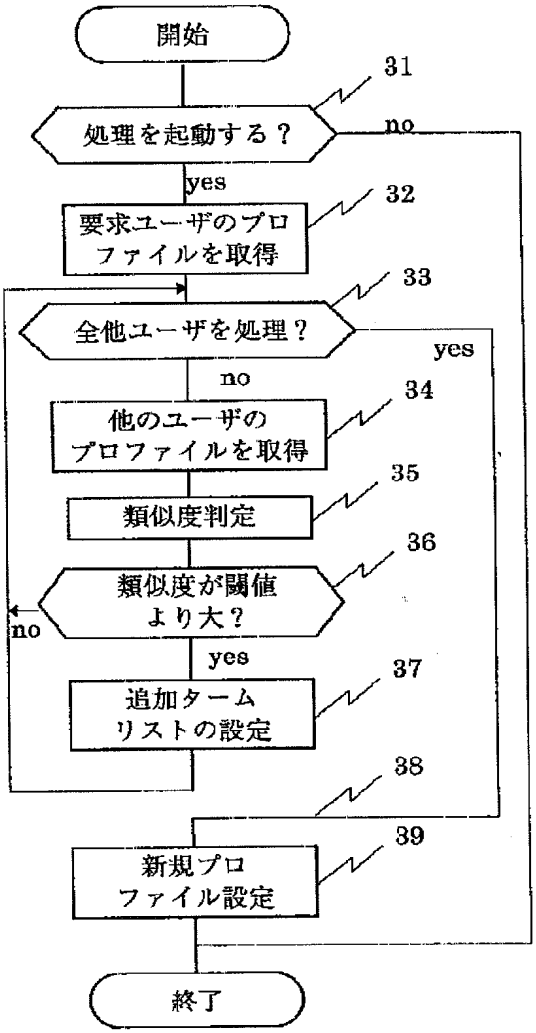
【図１２】

図１２


#3043 暗号を利用した… 表示
#4056 セキュリティを… 表示
#5026 インターネットでは… 表示
#5036 個人認証を… 表示
…

【図１４】

図１４



フロントページの続き

(72)発明者　梶　博行
東京都国分寺市東恋ケ窪一丁目280番地
株式会社日立製作所中央研究所内

Ｆターム(参考)　5B075 KK07 ND03 NK02 NK35 NR02
NR12 PR06 PR08 QM08